LYVE DATA LABS | SUSE We adapt. You succeed. SEAGATE

TECHNOLOGY PAPER

A Storage Evolution to Meet the Data Explosion

SUSE SES and Seagate Ceph Reference Architecture

INTRODUCTION

Many of our customers integrate Seagate® and SUSE products into their own solutions. There are many options when it comes to storage. The purpose of this document is to showcase an integrated and tested solution by Seagate and SUSE based on Ceph and Exos® E 4U106.

INTENDED AUDIENCE

This reference architecture (RA) is targeted at executives, managers, IT professionals, and administrators who deploy software-defined storage solutions within their data centers and make the different storage, services accessible as part of a manageable private cloud. By following this document, as well as those referenced herein, the audience should have a full view of the storage architecture, deployment, and administrative tasks, with a specific set of recommendations for deployment of the hardware and storage platform.

TODAY'S INDUSTRY FACES A MAJOR STORAGE CHALLENGE

Customers of all sizes face a major storage challenge, which we can subdivide into three buckets:

1. Cost: $/petabyte
2. Data size: The volume of data to be managed
3. Technical challenges: Data needs to be secure, available, and easily managed.

While the overall cost per terabyte of physical storage has gone down over the years, a data growth explosion, driven by the need to access and leverage new data sources (e.g., IoT [Internet of Things], edge sensors, videos, pictures, sensor data from autonomous vehicles, and drones) and the ability to *manage* new data types (e.g., unstructured or object data) has taken place.

These ever-increasing, let's call them *data lakes*, need different access methods: file, block, or object. Addressing these challenges with legacy storage solutions would require a number of specialized products (usually driven by access method) with traditional protection schemes (e.g., RAID). These specialized solutions struggle in particular when scaling from terabytes to petabytes at reasonable cost and performance levels. In addition, closed-source software and proprietary hardware leave customers in difficult situations when a product line is discontinued. This often requires a lengthy, costly, and disruptive complete replacement of EOL deployments.

SEAGATE BACKGROUND

Data in general is the lifeblood for businesses. The most prominent data-centric computing architectures are clouds, and being able to store unstructured data, images, and objects efficiently and cost-effectively is key. That's where Seagate hard disk drives (HDD) excel and are a perfect medium. No wonder the big public cloud architectures evolve around HDDs. Utilizing the latest highest-capacity drives to deliver the lowest cost per TB and creating scaleout object platforms like CEPH enable you to build PB building blocks for your private cloud.

HDDs have been used for nearly as long as modern computers have been in existence. HDDs are found in many common consumer products, such as notebook and desktop computers, gaming consoles, set-to-box DVRs, personal backup drives, and home network drives. HDDs are also enabling active data and archive data storage solutions for the enterprise and cloud service providers. For each usage category, specialized HDDs have been designed that are optimized for different characteristics, such as I/O performance, capacity, and cost. In aggregate, about 400 million drives are manufactured each year.

Seagate has been an industry-leading pioneer in the development of present-day drive technologies, from introducing the first Winchester technology 5¼-inch drives using longitudinal recording in the late 1970s to perpendicular recording in the mid 2000s to the latest innovation of heat assisted magnetic recording (HAMR), which will enable drive capacities of 20TB to 50TB by 2026. Seagate manufactures more drives worldwide than any other manufacturer[1] and is highly vertically integrated, designing and manufacturing all of the key components used in the product.

The **Seagate Exos E 4U106** storage system is one of the first in our portfolio to introduce our new modular approach to system design—an innovation that delivers a versatile architecture that's built to grow.

This white paper describes the design and implementation considerations Seagate has made in developing a Ceph-based solution on white box servers specifically sized to maximize cost-performance efficiency. While this implementation is specific to Seagate's choices of servers and storage, it is generally applicable to other types as well.

SUSE BACKGROUND

SUSE, the world's largest independent open source company, powers digital transformation with agile, enterprise-grade, open source solutions, from edge to core to cloud. With over 25 years of collaboration with partners, communities, and customers, SUSE delivers and supports enterprise-grade Linux, software-defined infrastructure, and application delivery solutions to create, deploy, and manage workloads anywhere—on premise, hybrid, and multi-cloud—with exceptional service, value, and flexibility.

SUSE Enterprise Storage delivers a highly scalable, resilient, self-healing storage software system. It is designed for private cloud environments ranging from hundreds of terabytes to exabytes. This software-defined storage product can reduce IT costs by optimizing data placement with the ability to automatically move data between tiers and leveraging industry-standard servers to present a unified storage servicing block, file, and object protocols. Having storage that can meet the current needs and requirements of the data center while supporting topologies and protocols demanded by new web-scale applications enables administrators to support the ever-increasing storage requirements of the enterprise with ease.

1 IDC, Worldwide 3Q18 HDD Shipment Results and Four-Quarter Forecast Update, John Rydning

A PROMISING SOLUTION

SUSE Enterprise Storage is based on Ceph, an open-source software-defined storage solution that enables transformation of the traditional storage infrastructure by providing a unified platform where structured and unstructured data can coexist and be accessed as file, block, or object, depending on the application requirements. 83% of Ceph users are either satisfied or extremely satisfied, as reported in a Ceph user **survey** in 2018. Ceph addresses all major storage challenges mentioned above: cost, scalability, is open source, and has the right technical features, such as data security and high availability. The combination of Seagate's Exos E 4U106, SUSE Enterprise Storage, and industry-standard servers furthermore promises to reduce cost for users while providing scalability to the exabyte level. That is critical to keeping up with future storage demands, which will grow worldwide 61% to 175 zettabytes by 2025, with as much of the data residing in the cloud as in data centers **(IDC Source)** With the sheer volume of data forecasted to be created over the coming years, businesses demand systems that are unified, distributed, reliable, highly performant, and, most importantly, massively scalable to the exabyte level and beyond. Ceph and the Exos E 4U106 storage system are a true solution for the world's growing data explosion, a space-conscious storage model that packs up to 1.6PB of raw storage capacity into a single chassis. Ceph's growth and adoption in the industry will continue at a lightning pace through the vibrant community of users who truly believe in the power of Ceph. Data generation is on an exponential growth curve, and we need to evolve storage to accommodate the explosive volume.

WHY CEPH?

Ceph is being broadly adopted for enterprise storage needs. It is tightly integrated into OpenStack, which is ideal for private cloud requirements for backup and archive use cases. It found its adoption in many industries like telco, AI and IoT. It can instantly provision hundreds of virtual machines from a single snapshot and build fully supported private clouds on standard hardware.

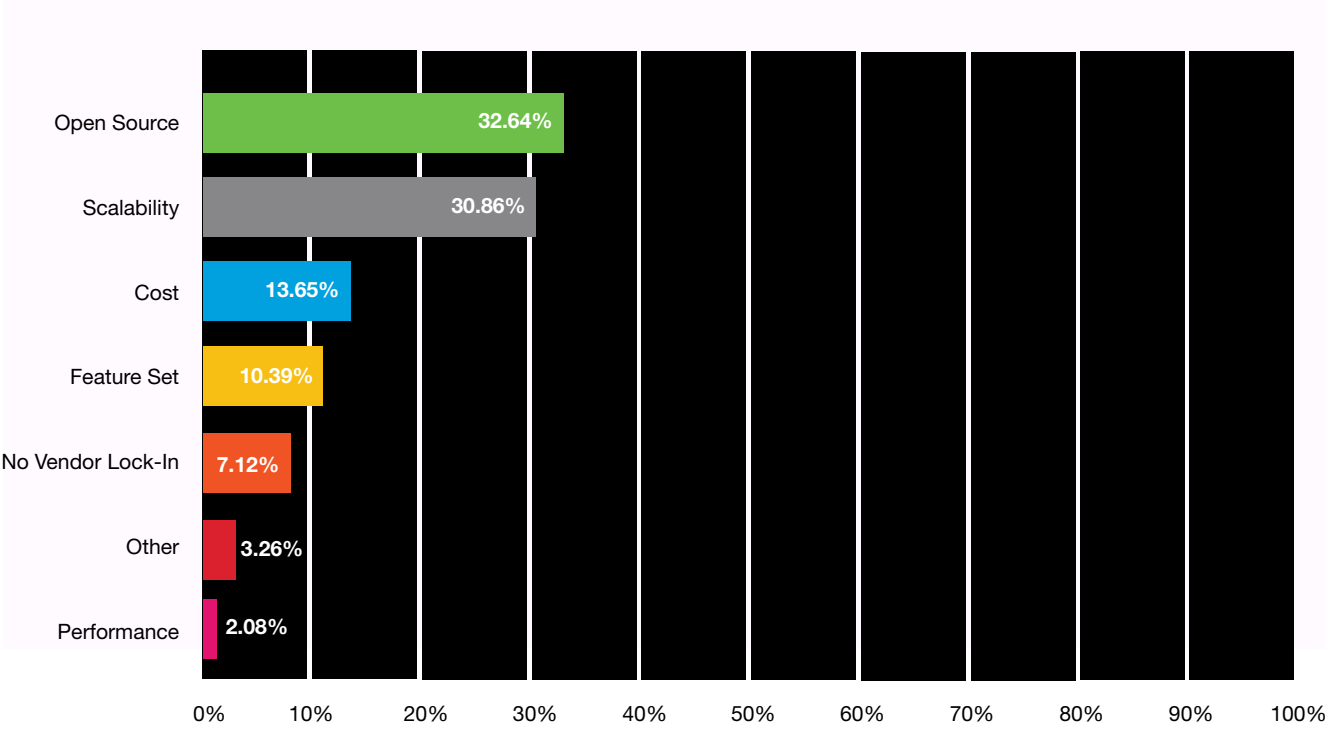
Let's talk about why companies would choose to use Ceph.

- **Highly Scalable—**Ceph is highly scalable. With Ceph, your storage cluster can start out as small as a single server for a mini proof of concept (POC), or you can grow the cluster when you want with very little effort by simply adding more servers and drives. You can scale over time in sync with your needs and budget without ever hitting a reasonable maximum storage limit.
- **Object, Block, and File storage—**Ceph can provide all of these with one cluster.
- **Scalable Performance—**Ceph has no "centralized" registry for data to flow through, so you don't get that bottleneck of data when you add more storage, clients, etc. In fact, Ceph can automatically balance the system to achieve the best utilization of storage resources across the cluster. Ceph's native protocols know where the data is and are able to connect clients directly to OSD processes, thus enabling the aggregate performance.
- **Flexibility—**You're not limited to homogeneous hardware nodes when adding more storage capacity into the cluster. For example, you can add an Exos E 4U106 to a Ceph cluster, or even add an all-flash array, such as the AFA 5005, to what was an all-spinning disk cluster. You can also add any other type of server to your cluster, which allows you to make use of legacy storage servers that you may own.
- **Self-Healing and Self-Managed—**If Ceph recognizes that a node goes down, it will start replicating your data to a new location in the background so it's always stored redundantly.

This list describes Ceph features and their value for your organization:

Feature	Means	Final Benefit
Open source	No vendor lock-in	Lower cost
Software-defined	Different hardware for different workloads	Broader user cases, higher efficiency
	Use commodity hardware	Lower cost, easier to evaluate
Scale-out	Manage many nodes as one system	Easier to manage = lower operational cost
	Distributed capacity	Multi-PB capacity for object and cloud
	Distributed performance	Good performance from low-cost servers
Block + object	Store more types of data	Broader use cases
Enterprise features	Data protection	Won't lose valuable data
	Self-healing	Higher availability, easier management
	Data efficiency	Lower cost
	Caching/tiering	Higher performance at lower cost

In a 2018 user survey, users of Ceph reported that it is most important to them that Ceph is open source, followed very closely by it's scalability capability. If you need scalability, you are in good hands.



GENERAL CEPH ARCHITECTURE

HOW CAN CEPH SCALE SO EASILY?

Objects are not tied to a physical path in Ceph, making objects flexible and location-independent. This enables Ceph to scale linearly from the petabyte level to an exabyte level. A key operational function is the CRUSH (controlled replication under scalable hashing) algorithm. Instead of performing a lookup in the metadata table (as done in traditional storage systems for every client request), the CRUSH algorithm enables the client to independently compute where data should be written to or read from. By deriving this metadata dynamically, therefore, there is no need to manage a centralized table, thus removing any potential bottleneck when scaling capacity up. It also enables Ceph's self-management and self-healing features.

CEPH COMPONENTS

Many find Ceph very complex due to its flexibility. However most of the complexity revolves around it's core layer, the RADOS (reliable autonomic distributed object store) layer, which is the heart of Ceph. All other components help to manage, keep track of metadata, or allow access to data. Let's look into each component:

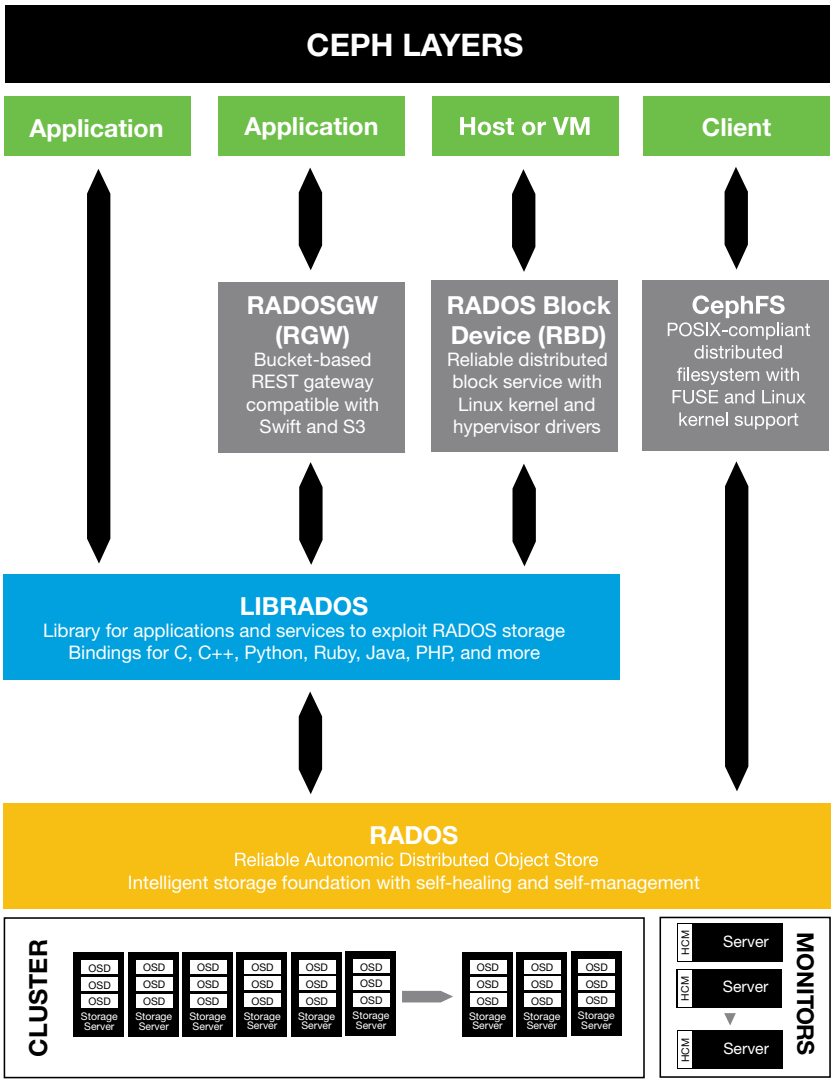
MON (Ceph monitor): Ceph monitors track all the metadata needed to run Ceph. They monitor the health of the entire cluster by keeping a map of the cluster state. They maintain a map of information for each Ceph component (OSD map, MON map, PG map, and CRUSH map). All the cluster nodes report to monitor nodes and share information about every change in their state.

Ceph manager: The Ceph manager daemon (ceph-mgr) runs alongside the monitor daemons to provide additional monitoring and interfaces to external monitoring and management systems.

Ceph OSD (object storage device): Ceph OSDs store the actual user data. Every OSD daemon is usually tied to one physical disk in the cluster.

RADOS (reliable autonomic distributed object store): It is the heart of the Ceph storage cluster. Everything in Ceph is stored in the form of objects irrespective of their data types. The RADOS layer keeps that data consistent through data replication, failure detection, recovery, data migration, and rebalancing across cluster nodes.

RBDs (RADOS block device): RBD is the native Ceph block device. It enables persistent block storage protocols including critical features such as thin-provisioning and being resizable.



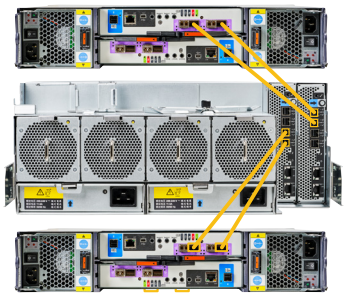
RGW (RADOS gateway interface): RGW provides the object storage service through a RESTful API with interfaces that are compatible with Amazon S3.

CephFS: The Ceph filesystem provides a POSIX-compliant filesystem service to users.

MDS (Ceph metadata server): The MDS stores metadata for the CephFS filesystem to keep track of file hierarchy. It does not serve data directly to clients.

Librados: The librados library allows access to the RADOS layer through support of many programming languages, such as Ruby, Python, PHP, etc. This provides a native interface for the Ceph storage cluster (RADOS) and is the base component for all native Ceph services, such as RBD, RGW, and CephFS.

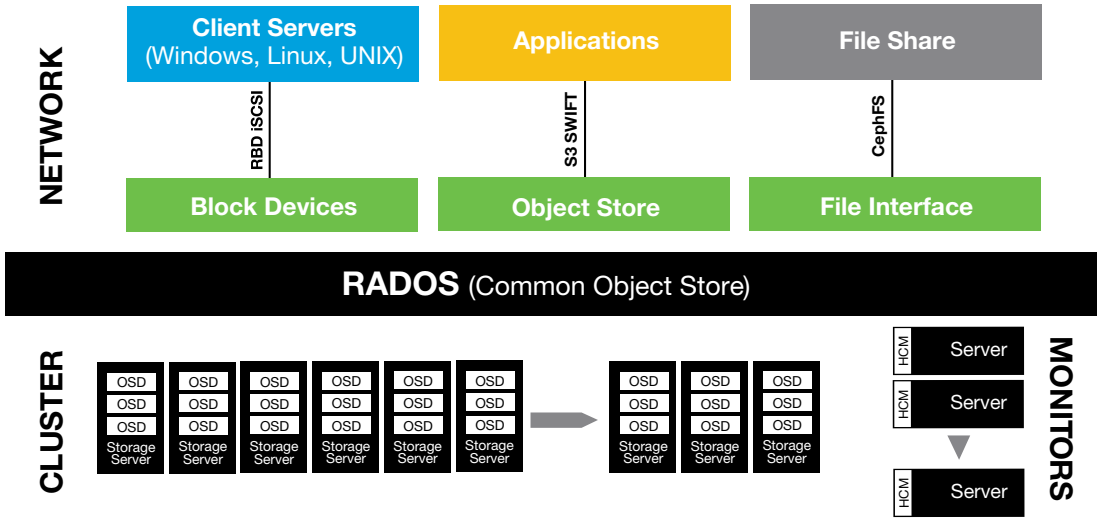
REFERENCE ARCHITECTURE OVERVIEW



The Seagate-SUSE Enterprise Storage reference architecture (RA) is based on a **building block** consisting of two compute nodes and one Exos E 4U106 JBOD, which serves 53 disks each to the two servers.

The full setup is made of three such building blocks to form a 6-node Ceph cluster with up to ~4.5PB, hence it is based on six storage servers using Intel® Servers with Intel Xeon® Gold 6140 processors, three monitor servers, and one admin server. This combination provides the high CPU performance required for a performance-optimized Ceph cluster and yields an open, cost-optimized platform. This platform can be utilized as an effective building block for implementing a multi-petabyte cloud storage infrastructure.

Below is a logical architecture picture describing our setup:



SOLUTION CONSIDERATIONS

SIZING CONSIDERATIONS

The following section provides insight and understanding of the performance and sizing considerations typical for Ceph and its components. Building a Ceph cluster is a careful balancing act between storage, network speeds and CPU requirements while also taking density and budget into consideration.

OSD JOURNAL PERFORMANCE

BlueStore is a new storage back end for Ceph. It boasts better performance (roughly 2× for writes), full data checksumming, and optional built-in compression. It is the new default storage back end for Ceph OSDs and provides a huge advantage in terms of robustness and functionality.

SERVER SIZING

OSD Memory needs

By default the OSD process assigns cache based on the device class. If the device is an HDD, the cache assignment is 2GB, and for SSDs is 3GB. This is a user-configurable variable. Each OSD also needs 2GB of RAM for the base processes to operate. 16GB is a starting point that provides some buffer.

A good rule of thumb to calculate the needed memory is 16GB + (2+OSD cache) * OSD count.

In this case: 16GB + (2+2) * 53) = 228GB RAM.

CPU

MON nodes: four cores

RGW nodes: eight cores

MDS nodes: four-eight cores at a higher clock rate

OSD nodes: one core-GHz per HDD, two per SSD, five per NVMe™

Real world example: OSD nodes with dual E5-2680 v2 Xeon processors driving 10 OSDs of 3-4TB LFF rotational drives each. These 10 core, 2.8GHz CPUs were mostly idle. Selecting E5-2630 v2 Xeons instead would have saved ~$2600 per system (list price) and mitigated C-state issues.

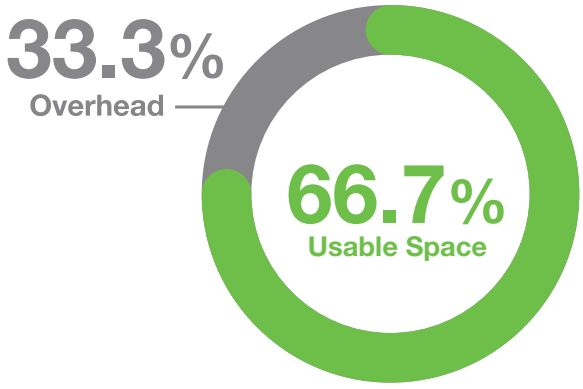
Real-world example: OSD nodes with dual E5-2683 v4 Xeon processors driving 24 OSDs of 1.6TB to 3.8TB SSD drives each. These 16-core, 2.1GHz CPUs are 10% to 20% utilized under load. The E5-2630 with 10 cores at 2.2GHz would save $2200 off the list price of each server and still provide sufficient cycles.

The latest SUSE recommendation is 1 × 2Ghz thread per spinning OSD.

STORAGE SIZING

For our setup, we will be using erasure coding 4+2, three fully loaded Seagate Exos E 4U106 systems, which will equal 2.1PB of usable space. For production, it is recommended to have +3 erasure coding setup. For test and performance tests, usually a 4+2 gives good indications.

No. of Drives	318	drives
Drive Capacity	10	TB
No. of Data Chunks	4	
No. of Coding Chunks	2	
Useable Space	2.1	PB
RAW	3.18	PB
Useable Space in %	66.67	%
Overhead in %	33.33	%



WHAT ABOUT SECURITY—ENCRYPTION?

Usually there are thousands of clients, and the data for each client is distributed among as many as hundreds of OSD devices. This mixing of data presents a challenge to someone trying to gain access to the data, even if they could gain physical access to your servers. However, when data is truly sensitive, requirements might dictate stronger protection in the form of encryption. There are two options to the user: A software-based encryption by using *--dmcrypt* switch when using Ceph, or a more efficient hardware solution, such as Seagate Secure™ Self-Encrypting Drives (SED) that implement encryption of data at rest and can be used transparently by Ceph.

SYSTEM ARCHITECTURE OVERVIEW

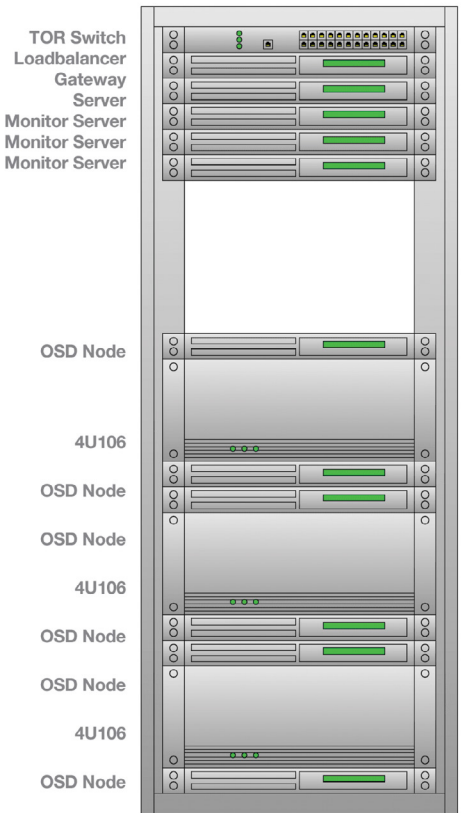
In this section of the paper, we will look at various configuration options for Ceph to optimize performance.

OPERATING SYSTEM SOFTWARE

All six server systems have SUSE Enterprise Storage 6 loaded and run on SUSE Linux Enterprise Server 15 SP1.

PHYSICAL LAYOUT

CEPH RACK

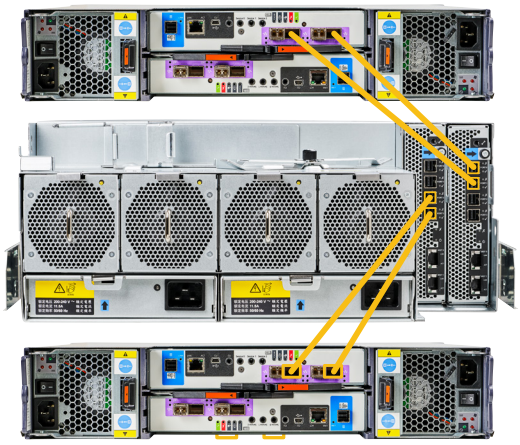


HARDWARE INFRASTRUCTURE

Storage system

The Seagate enterprise-class storage solution is being built leveraging a Seagate Exos E 4U106 high-density enclosure built with a next-generation operating system SUSE Linux enterprise server (SLES) that together deliver performance, capacity, and reliability. The Exos E 4U106 enclosure and firmware technology will enable businesses to store massive quantities of data in a high-availability enclosure that provides high-performance access to data. What's more, Seagate's operating system not only powers the high density, but dramatically improves the system's performance and reliability.

Cable infrastructure for each Exos E 4U106 (we use 1 HBA / 2 SAS ports)

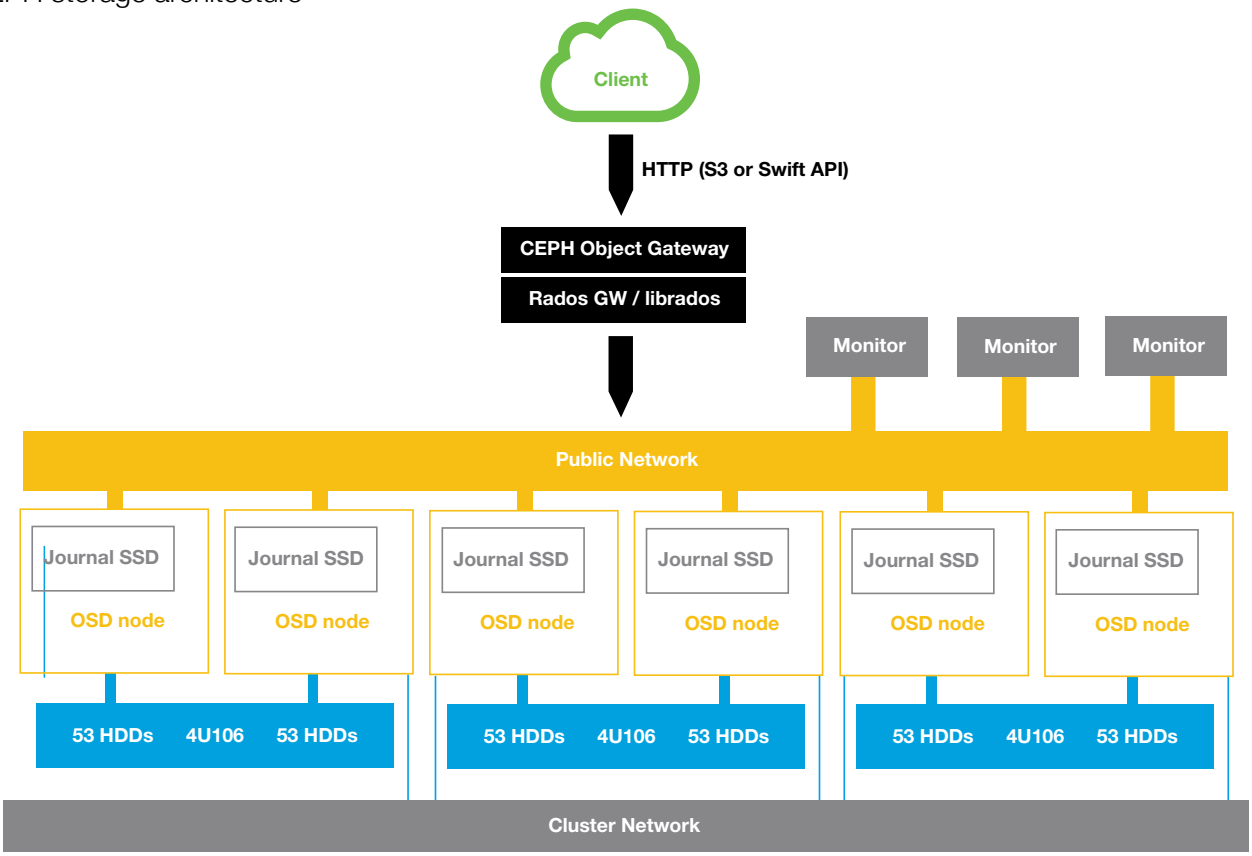


SERVER CONFIGURATION

Ceph Component	6 Storage Node (OSD)	3 Monitor Nodes
Platform	Intel Server	Intel Server
CPU	Xeon E5 6140	
Memory	192GB	
Network	100GB	
Storage	Each node has access to 53 disks (10TB) on Exos E 4U106	

LOGICAL SETUP

CEPH storage architecture

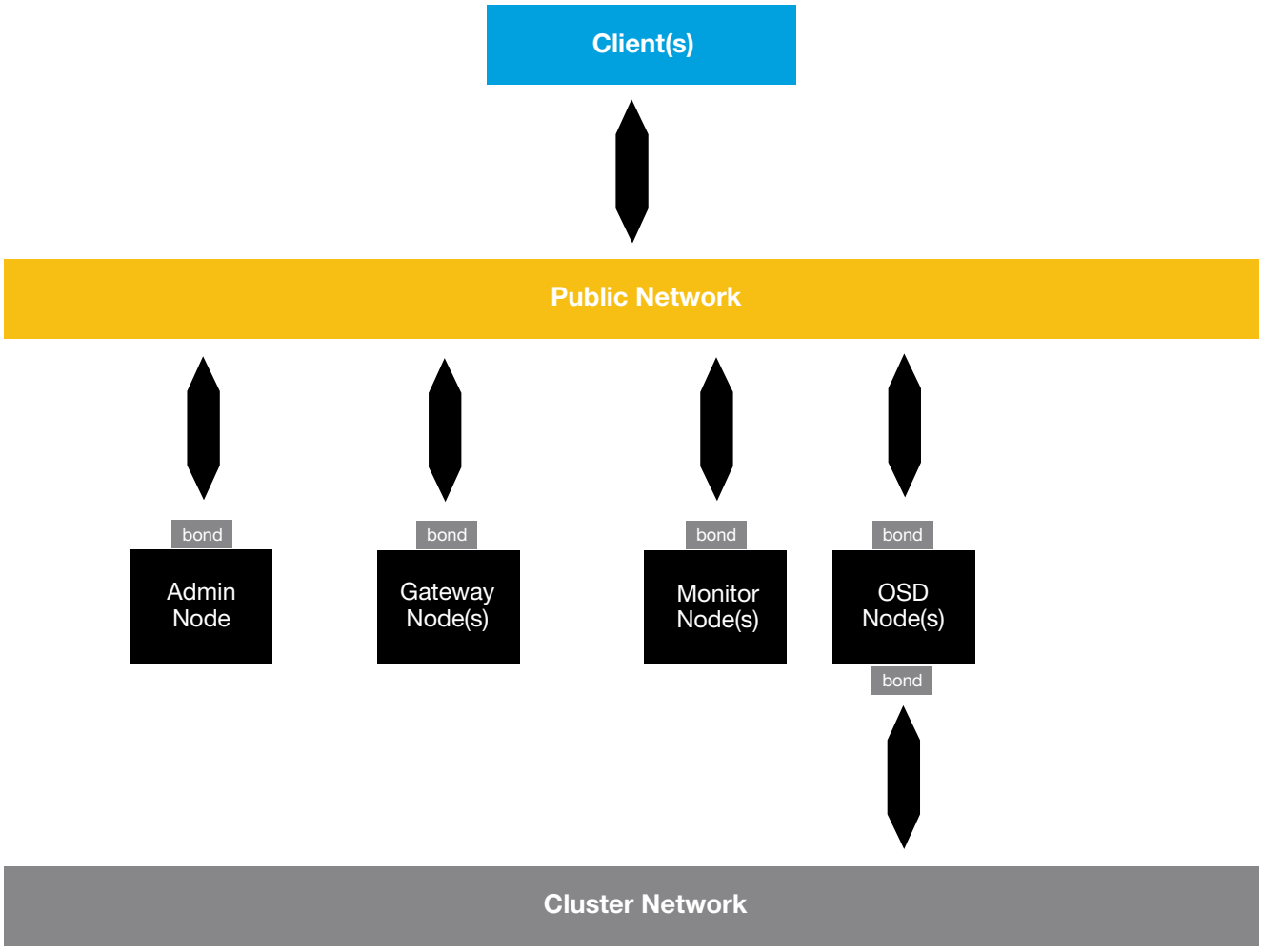


- One system is deployed as the administrative host server. The administration host is the Salt master and hosts the SUSE Enterprise Storage administration interface, CEPH Dashboard, which is the central management system that supports the cluster.
- Three systems are deployed as monitor (MON) nodes. Monitor nodes maintain information about the cluster health state, a map of the other monitor nodes and a CRUSH map. They also keep a history of changes performed to the cluster.
- Additional servers may be deployed as iSCSI gateway nodes. iSCSI is a storage area network (SAN) protocol that allows clients (called initiators) to send SCSI commands to SCSI storage devices (targets) on remote servers. This protocol is utilized for block-based connectivity to environments such as Microsoft Windows®, VMware, and traditional UNIX. These systems may be scaled horizontally through client usage of multi-path technology.
- The RADOS gateway provides S3 and Swift based access methods to the cluster. These nodes are generally situated behind a load balancer infrastructure to provide redundancy and scalability. It is important to note that the load generated by the RADOS gateway can consume a significant amount of compute and memory resources, making the minimum recommended configuration contain 6-8 CPU cores and 32GB of RAM.
- SUSE Enterprise Storage requires a minimum of four systems as storage nodes. The storage nodes contain individual storage devices that are each assigned an Object Storage Daemon (OSD). The OSD assigned to the device stores data and manages the data replication and rebalancing processes. OSDs also communicate with the monitor (MON) nodes and provide them with the state of the other OSDs.

NETWORKING DIAGRAM

From a networking perspective, Ceph should be built using two networks.

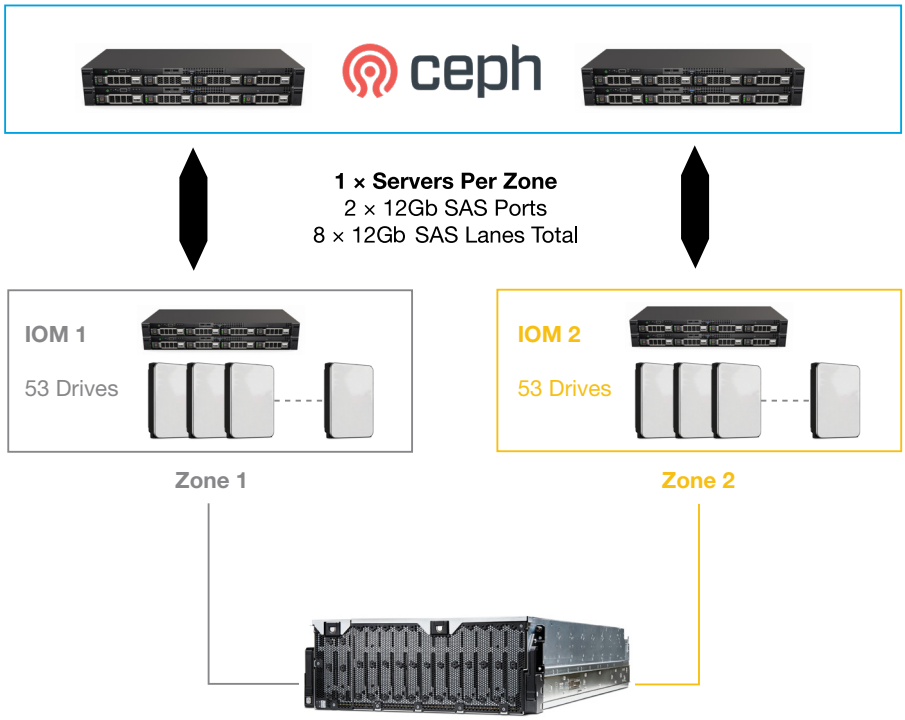
1. A public-facing network that clients can access storage from
2. A cluster network that is used for cluster internal communication to handle OSD heartbeat, object replication, and recovery traffic



OSD NODE SETUP

OSD nodes are a key component for Ceph. All OSD compute nodes need to be able to communicate with each other and can be installed using Ceph native methods. We will use two OSD nodes for each Exos E 4U106. Each 4U106 needs to be configured with a so-called *shared nothing* configuration. This will split the control into two sets of 53 disks. This will allow us to run 53 OSDs per OSD node and keep the CPU/memory footprint within economical boundaries and also the failure domain smaller.
As depicted below, each I/O module will be configured to run with one OSD node.

OSD Nodes: Two 53 OSD/Drive Zone Architecture



PLACEMENT GROUPS

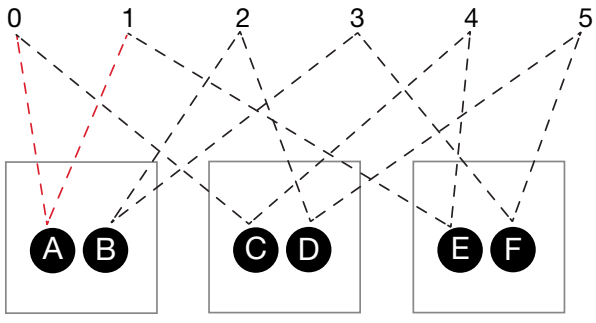
Placement groups (PG) are an internal implementation detail of how Ceph distributes data. You can allow the cluster to either make recommendations or automatically tune PGs based on how the cluster is used by enabling pg-autoscaling. Allowing the cluster to automatically scale PGs based on usage is the simplest approach, which we use in our setup.

If one wants to manually control PGs, following are some considerations.

In essence, PGs provide a mechanism to control:

1. The level of replication declustering, which means how many physical hard drives/machines a given volume is distributed over (which can be used for load distribution)
2. How many other drives/machines a given drive/machine shares a portion of its data with (declustering)
3. It's basically a replication policy
4. Scaling a cluster

This illustration shows a simple placement group (PG=6).



PERFORMANCE RESULTS

TEST PLAN

Performance test parameters

Number of COSBench clients: 1, 6, 12, (…), 120
Threads per COSBench client : 1, 8, 16, 32, 64, 128, 256
Object size: 512KB, 1MB, 4MB, 8MB
I/O pattern: 100% write, 100% read
Runtime of each test: 5 minutes
Average of 5 runs
Lightly tuned cluster

Baseline testing

1. Maximum speed

Performance testing:

1. 1 client testing (object size 512k, 1MB, 4MB, 8MB) throughput / no. of threads (1/8/16/32/64/128)
    a. Write performance tests with one client
2. 6 clients testing
    a. Write performance with six clients
    b. Disk performance (write, read, throughput)

20 clients testing

1. Write performance
2. Disk performance (write, read, throughput)

4+2 erasure code testing

1. Write test: 4+2 pool 1MB object
2. Write test: 4+2 pool 8MB object 120 COSBench clients
3. Write test: 4+2 pool 1, 4, 8, 16, 32, 64MB object 90 COSBench clients/9000 threads

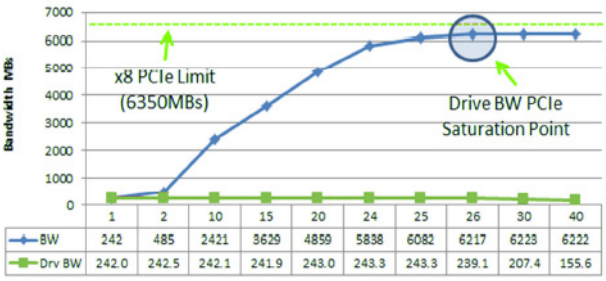
BASELINE TEST RESULTS

Maximum Exos E 4U106 speed

The maximum speed you can reach with an Exos E 4U106 is theoretically capped at the ×8 PCIe® speed limit. Practically, we get close to this as illustrated below.

Single ×8 SAS HBA to JBOD results - raw bandwidth

4U106 Single ×8 SAS to Right IOM 1M Seq RD



	1	2	10	15	20	24	25	26	30	40
BW	242	485	2421	3629	4859	5838	6082	6217	6223	6222
Drv BW	242.0	242.5	242.1	241.9	243.0	243.3	243.3	239.1	207.4	155.6

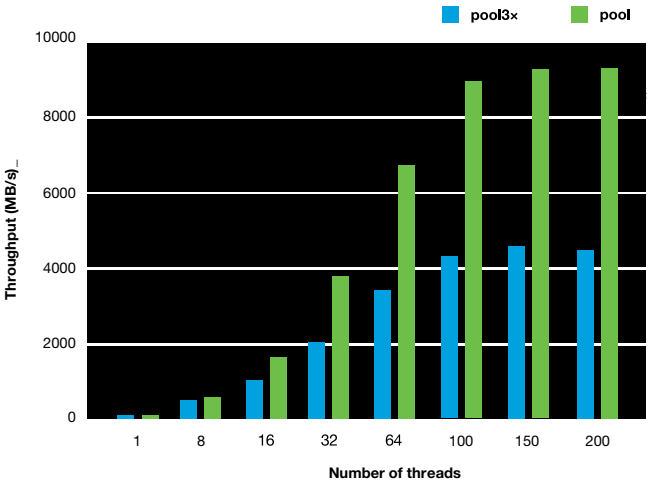
PERFORMANCE TEST RESULTS

It should be noted that only light tuning of the test environment has been performed, leaving significant potential for increased performance with further tuning. The replicated pool results are intended to act as a baseline for comparison with the erasure-coded pool results.

COSBench was used to run the performance tests. COSBench can be **downloaded here**.

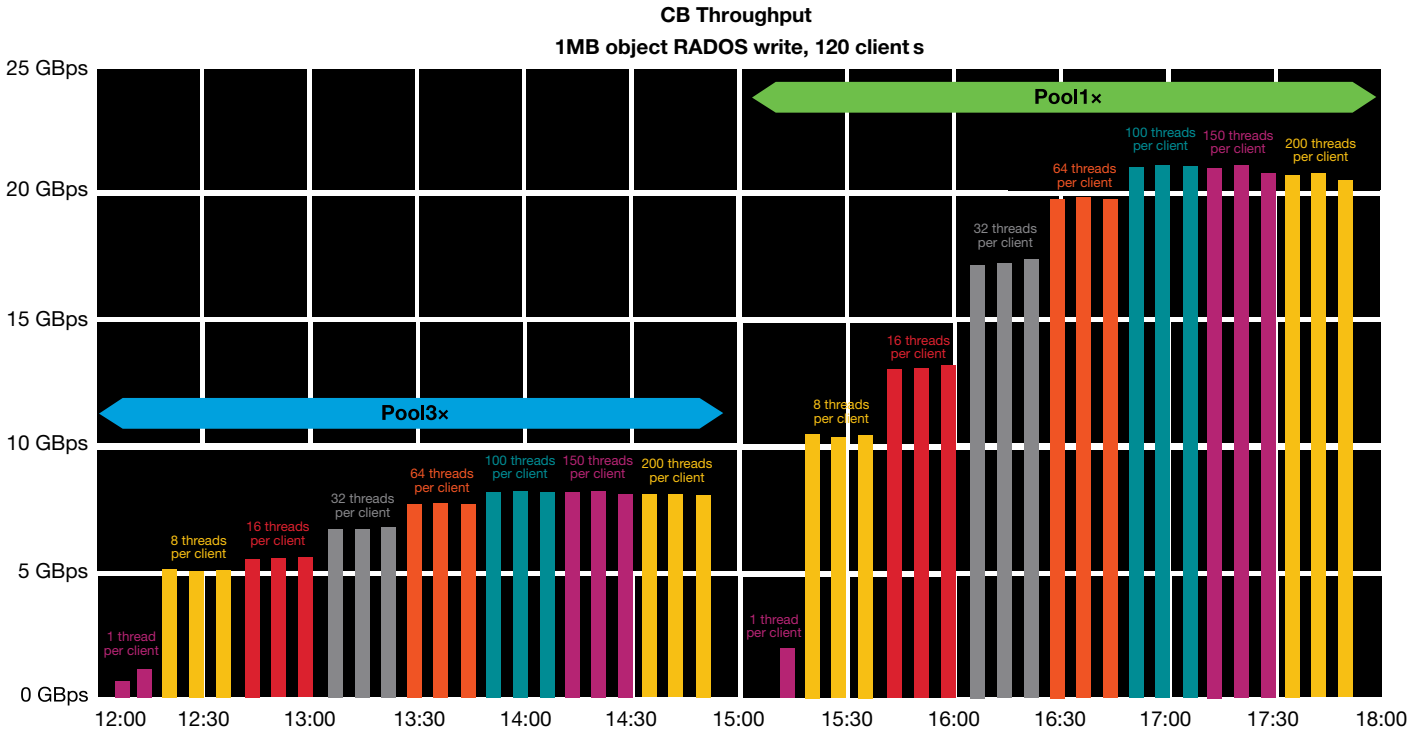
Write performance: 6 clients, object size 1MB

The difference in performance between 1× and 3× replicas clearly illustrates the overhead required for replication. The performance scaling shows a strong relationship to the number of threads.

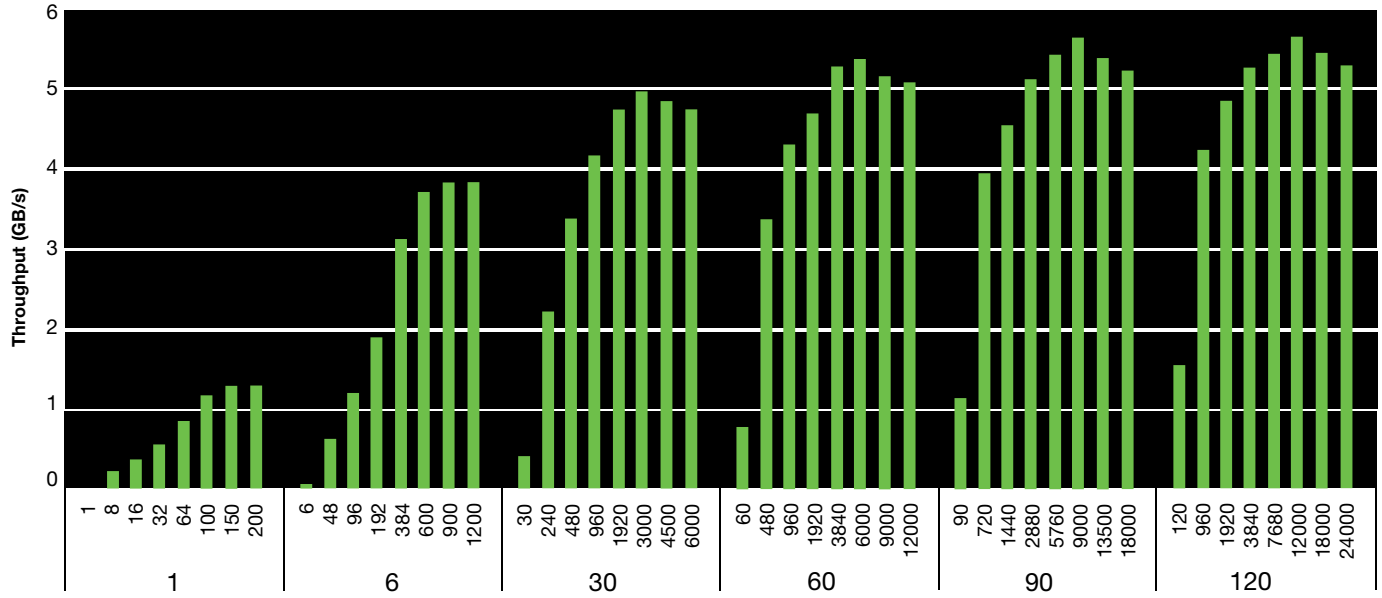


Write throughput: 1MB object, 120 clients

The difference between 3× replica and 1× replica illustrates clearly the overhead needed to perform back-end replication.

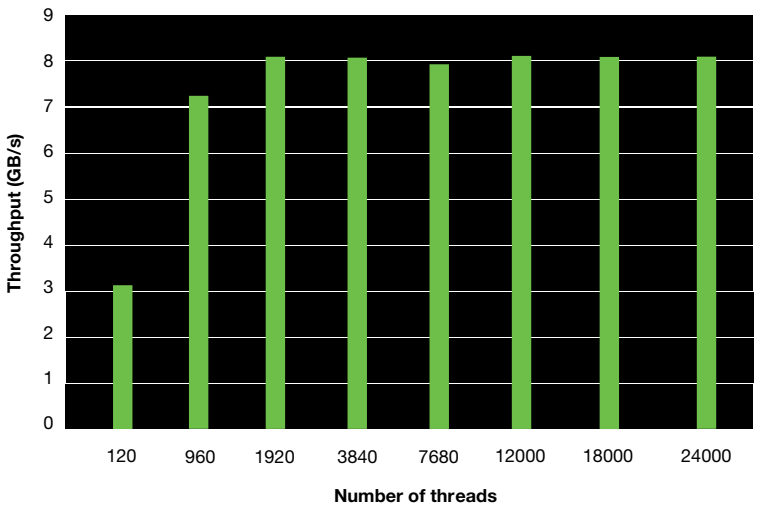


4+2 erasure code write performance, 1MB object size

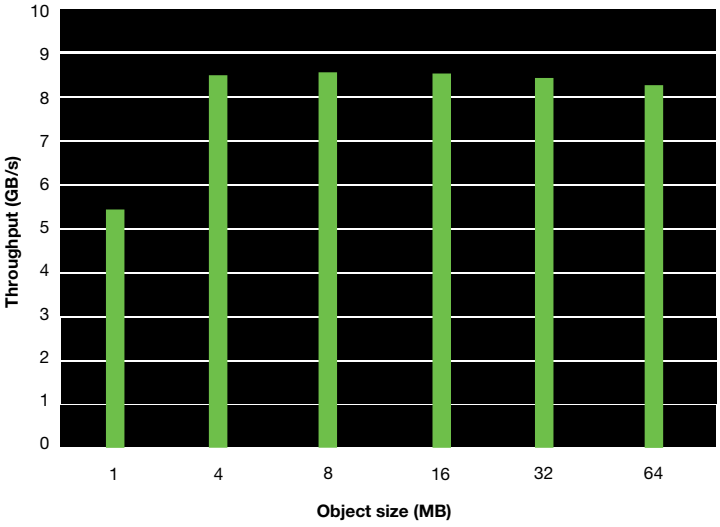


On the following performance measurements, it can be noted that the 25Gb/s link of the three RADOS gateways were rapidly saturated.

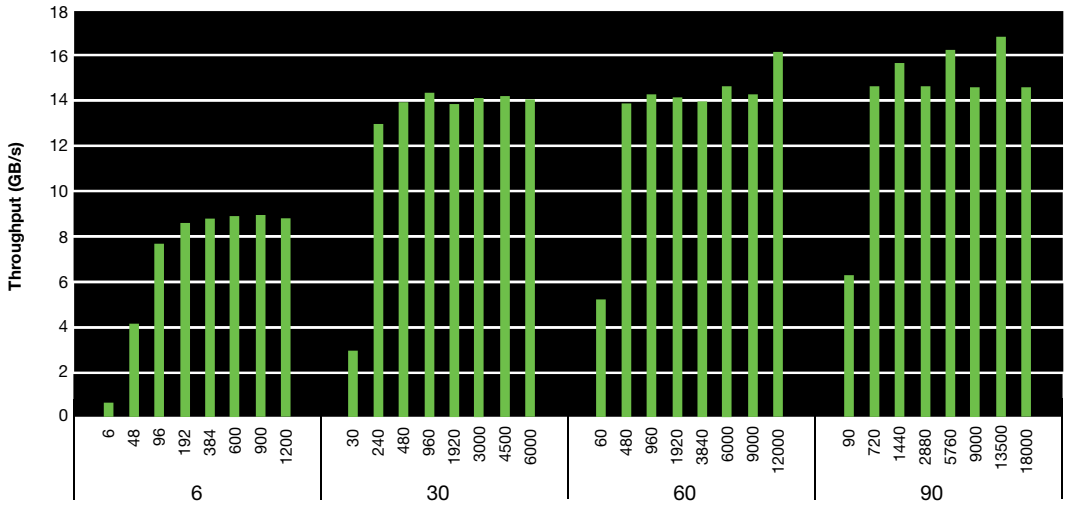
4+2 erasure code write performance, 8MB object size, 120 COSBench clients



4+2 erasure code write performance, 90 clients, 9000 threads



4+2 erasure code read performance, 8MB object size



Summary of results

Overall, the following table depicts the test results for 1×, 3× and erasure coded pools.

	Max Performance	Max Performance per Disk
3×-write	8.3GB/s	26.8MB/s
1×-write	22.15GB/s	71.3MB/s
(1×/3×)-read	31.07GB/s	100MB/s
4+2-write	8.7GB/s	28MB/s
4+2-read	16.7GB/s	54MB/s

HARDWARE COMPONENT LIST

The hardware used for the implementation of the reference architecture:

STORAGE HARDWARE

HW - Part	Type	Comment
3	Seagate Exos E 4U106 Exos systems	JBOD with redundant I/O modules
3*106 = 318	Seagate Exos X10 10TB HDDs	Storage hard drives

SERVER HARDWARE

HW - Part	Type	Comment
20	Intel Xeon gold 6140, 18C, 2.3GHz, 24.75MB cache, DDR4 up to 2666MHz, 140W TDP, socket FC-LGA14	CPU's Fully loaded servers (OSD, monitoring, admin)
6	1U Intel server system R2208WF0ZSR	Server enclosures OSD
4	2U Intel server R2208WF0ZSR 2U RM (2×) CPU slots 24×DIMMs 8×2.5" HS bays 1×1100W	Server enclosure Admin Node Monitoring nodes
	mSAS-HD cable kit AXXCBL650HDHRT	
96	Crucial 16GB DDR4-2666 RDIMM 16GB DDR4-2666 RDIMM 1.2V CL19	Memory / RAM
10	ConnectX-5 EN network interface card, 100GbE dual-port QSFP28, PCIe3.0 ×16, tall bracket, ROHS R6	Networking for each server
10	Remote Management Module 4 Lite 2	For each server
10	LSI 9300-8E SGL 8PORT 12GB/s]	SAS HBA for each OSD

NETWORKING HARDWARE

HW - Part	Type	Comment
1	GE RJ45 1U open Ethernet switch with ONIE, 48-port GE RJ45 port + 4×10G SFP+, 2 power supplies (AC), Integrated ARM A9 CPU, short depth, fixed fans, P2C airflow, rail kit	Top of rack switch
2	100GbE 1U open Ethernet switch with Cumulus Linux, 16 QSFP28 ports, 2 power supplies (AC), ×86 CPU, short depth, P2C airflow, rail kit must be purchased separately	100GB Switch
	Kit, cable, support	